CogAgent: A Visual Language Model for GUI Agents

Year: 2023 **|** *Citations: 560* **|** *Authors: Wenyi Hong, Weihan Wang, Qingsong Lv*

Abstract

People are spending an enormous amount of time on dig-ital devices through graphical user interfaces (GUIs), e.g., computer or smartphone screens. Large language models (LLMs) such as ChatGPT can assist people in tasks like writing emails, but struggle to understand and interact with GUIs, thus limiting their potential to increase automation levels. In this paper, we introduce CogAgent, an 18-billion-parameter visual language model (VLM) specializing in GUI understanding and navigation. By utilizing both low-resolution and high-resolution image encoders, CogA-gent supports input at a resolution of1120 × 1120, enabling it to recognize tiny page elements and text. As a general-ist visual language model, CogAgent achieves the state of the art on five text-rich and four general VQA benchmarks, including VQAv2, OK- VQA, Text- Vqa, St- Vqa, ChartQA, infoVQA, DocVQA, MM-Vet, and POPE. CogAgent, using only screenshots as input, outperforms LLM-based methods that consume extracted HTML text on both PC and Android GUI navigation tasks-Mind2Web and AITW, ad-vancing the state of the art. The model and codes are available at https://github.com/THUDM/CogVLM.